DOCUMENT RESUME

ED 359 243 TM 020 012

AUTHOR Shen, Linjun
TITLE Constructing a Measure for Longitudinal Medical Achievement Studies by the Rasch Model One-Step Equating.
PUB DATE Apr 93
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993). Figures 3-10 contain small print of marginal legibility.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Equated Scores; Estimation (Mathematics); Higher Education; Item Response Theory; *Knowledge Level; Licensing Examinations (Professions); *Longitudinal Studies; Mathematical Models; Medical Education; *Medical Students; Osteopathy; Rating Scales; *Test Construction
IDENTIFIERS Ability Estimates; BIGSTEPS Computer Program; Missing Data; National Board of Osteopathic Medical Examiners; *Rasch Model

ABSTRACT

As part of a longitudinal study of the growth of general medical knowledge among osteopathic medical students, a simple, convenient, and accurate vertical equating method was developed for constructing a scale for medical achievement. It was believed that Parts 1, 2, and 3 of the National Board of Osteopathic Medical Examiners' (NBOME) examination together represented the basic concepts and principles of the entire medical sciences, so all 3 (comprising 9 examinations in all) were combined in a linking chain into a large examination representing 2,814 items and 5,168 persons, although there was substantial missing data. The Rasch model measurement program BIGSTEPS was used to calibrate this huge examination. Despite the large amount of missing data, the program converged smoothly. Distributions of person ability were not affected by the equating. Results suggest that Rasch measurement one-step equating is a valid, efficient, and accurate way to construct a measure for longitudinal medical achievement studies. The one-step technique accomplished by BIGSTEPS provided good person ability estimates on the whole examination, and consistent difficulty estimates for items. Ten figures and two tables present details of the analyses. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LINJUN SHEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED359243

Constructing a Measure for Longitudinal Medical Achievement Studies by the Rasch Model One-Step Equating

Linjun Shen

National Board of Osteopathic Medical Examiners

Paper presented at the Annual Meeting of the American Educational Research Association
Atlanta, April 1993

TM020012

Constructing a Measure for Longitudinal Medical Achievement Studies by the Rasch Model One-Step Equating

Linjun Shen

This analysis is part of a longitudinal study of the growth of general medical knowledge of osteopathic medical students in the 1987 cohort. The purpose of this paper is to present a simple, convenient, and accurate vertical equating method for constructing a measurement scale for medical achievement.

Although it is comprised of three relatively independent phases: preclinical education, clinical education, and residency, medical education is a continuum. Three different phases share common educational objectives. Therefore, assessing the change of desired cognitive traits is naturally an effective approach to evaluate the function and structure of medical education. Yet, partially due to the difficulties of constructing a longitudinal measurement scale, one of the major methodological deficiencies of research on medical education is the lack of longitudinal studies (McGuire, 1993; Gonnella, et al., 1993).

To study the growth of medical achievement, a valid measurement scale must (1) represent the entire medical curriculum (2) have a high psychometric comparability so that measures taken at different time points during the medical program will have a same qualitative and quantitative explanation. In the medical education literature, among others, there were three major experiments in making longitudinal scales of medical achievement: the "Minitest" of the NBME in the late 1960s' (Levit, 1967), the Quarterly Profile Examination (QPE) at the University of Missouri, Kansas City Medical School in the late 1970s' (Willoughby et al., 1978), and the Progress Test at the Maastricht Medical School in the Netherlands since early 1980s' (Verwijnen et al., 1990).

The NBME minitest was a one-day, 360 multiple-choice question examination designed for individual schools to evaluate their curricula. Items were drawn from 12 subject matter areas of Part I and Part II. The test was given on an annual basis in an attempt to track the learning process. QPE was developed to evaluate medical students' acquisition and retention of medical knowledge during their six-year program. Each QPE contained 400 items selected from a 10,000 item bank covering all clinical disciplines and some basic science disciplines. Each student took

the QPE four times a year. The Progress Test consisted of about 250 true-false items covering medicine as a whole. Items were selected from an IRT calibrated large item bank. The Progress Test was administered four times a year to the entire student body at the same time. All these studies attempted to keep track of growth by making tests cover the entire medicine and administering the tests on yearly or even quarterly basis.

One common feature is that all these tests had to maintain a substantial portion of items which were irrelevant to students' real achievement level. Several problems result. Firstly, there were not enough items targeting current achievement; secondly, students might not be well motivated by being forced to answer substantial amount of items which were beyond or below their actual achievement level; thirdly, construct validity could be threatened because students' readiness for items at the same achievement level were different when they progressed along the program.

To avoid these problems, this study only used exams which were designed to measure students' current educational level. The advanced Rasch measurement one-step equating technique made this design feasible (Lee, 1992).

Method

The 1987 cohort was measured by the National Board of Osteopathic Medical Examiners' (NBOME) June 1989 Part I (A891), March 1991 Part II (B911), February 1992 Part III (C921) at the end of preclinical, clinical and resident education respectively. If these three exams were on a common measurement scale, the scores of the three examinations would reflect the growth of students' general medical knowledge. However, Part I, II, and III examinations were traditionally constructed and analyzed as different examinations. Scores of three exams were not compatible. Therefore, the task for this study was to equate A891, B911, and C921.

Design

Though the NBOME Part exams did not have common items across Parts, one examination (S912) which certified beginning practitioners contained items from all three Parts. Indirect connections between three target exams and S912 were found. Figure 1 demonstrates how A891, B911, and C921 were eventually connected to S912.

Believing the NBOME Part I, II, and III together represented the basic concepts and principles of the entire medical sciences, this study combined all exams in the linking chain into a large exam, and assumed that this large exam defined the general medical knowledge. Figure 2 shows the data structure of this huge exam. Nine exams were appended to one another according to the item overlapping structure. Item sequences for each exam were reordered so that the responses for the same items in different exams were in the same columns. Consequently, 2814 items and 5168 persons were included. However, 85% of the artificial exam were missing data.

Subjects

All students in A891, B911, C921, and S912 were included in the equating, while students in other non-target exams were randomly selected. To make the equating efficient, students from non-target exams all had person fit statistics less than 2.00.

Instruments

This equating involved nine NBOME examinations. All exams were certification examinations for osteopathic medicine. There were three Part I exams (A891, A901, A921), two Part II exams (B871, B911), three Part III exams (C871, C881, C921), and one non-Part certification exam (S912). Part I and II examinations covered basic sciences and clinical sciences respectively. Part III examinations covered the same disciplines as Part II but all questions came from clinical practice emphasizing problems with high impact and high frequency. S912 covered all biomedical disciplines taught in medical school but emphasized clinically relevant knowledge. S912 was given to candidates who just finished their residency program. All exams in this study had a reliability coefficient of .90 or higher.

This study identifies all exams by four characters. The first character designates the Parts with A for Part I, B for Part II and C for Part III. The second and third characters are for the year of the exam; the last is for the administration. So, for example, B912 refers to the second administration of 1991 Part II exam.

Procedures

The latest Rasch measurement program BIGSTEPS was used to calibrate this huge exam. To obtain a good estimation, three convergence criteria were set prior to the calibration: (1) the

maximum logit change is .00 (2) the maximum raw score change is less than .00 (3) the maximum number of PROX is 300, the maximum number of UCON is 300. An 386-20 Mhz IBM-compatible computer with a math-coprocessor executed this program.

Analysis

Dimensionality The fundamental assumption of this vertical equating is that medical knowledge is an entity and basic biomedical sciences and clinical sciences are components of this integrated domain. However, when the task is to equate three types of medical certification examinations which focus different components of medicine and have different degree of clinical relevance, it is necessary to assess the dimensionality of the equated scale. This study assumes that if the equated scale measures something different from each of the local scales, items and persons' response patterns will differ. The fit statistics will reflect such differences. Therefore, this study compared both person and item mean square infits before and after equating. Consistency of the fit provides evidences of the unidimensionality of the bank scale.

Scale equity A crucial issue of vertical equating is the equity of scales before equating (local scale) and after equating (bank scale). Previous studies indicate that measurement scales have dramatic consequences on the outcomes of growth study (Becker & Forsyth, 1993; Schulz, Shen & Wright, 1990; Yen, 1986). To investigate the scale equity, this study assessed the person ability distributions before and after equating.

Sample indifference A unique feather of one-step equating is there are a large amount of missing data. Local scales and the bank scale were derived based on different groups of subjects. Hence sample invariance is of special importance to one-step equating. Based on the different involvement of subjects in bank scale calibration, Items in each exam could be grouped into several subsets. Specifically, S912 had four subsets: subset 1 contained 17 items shared by S912 and A921, subset 2 had 100 items shared by B871, subset 3 shared 125 items with C871 (part of them also shared by C881), subset 4 was 209 items which did not overlap any other exams. A891 had three subsets: subset 1 had 18 items shared with A921, subset 2 had 722 non-overlapped items, subset 3 was 142 items shared with A901. B911 had two subsets: subset 1 had 60 items with B871, subset 2 was 772 non-overlapped items. C921 had two subsets: subset 1 was 35 items with C881, subset 2 was 541 non-overlapped items. To assess if the scale

transformations were consistent across different sets of items within a same exam, this study computed $\overline{d}$, or the mean of $d_i=d_{iB}-d_{iL}$, for each item subset and total exam, where d_{iB} was the difficulty of item i on the bank scale and d_{iL} was the difficulty of item i on the local scale. $\overline{D}$ was the observed difference between $\overline{d}$ for the total exam and $\overline{d}$ for each item subset. Z_i was the standardized d_i. $\overline{Z}$ was the mean of Z_i. The square root of mean square residual of Z_i, or RMS_{Z_i}, was also computed

$$RMS_{Z_i} = \sqrt{\frac{\sum (Z_i-\overline{Z_i})^2}{N-1}}$$

where

$$Z_i=\frac{d_{iB}-d_{iL}-\overline{D}}{SE_{d_{iB}-d_{iL}}}$$

RMS_{Z_i} was the summary statistics of the scale conversion. A RMS_{Z_i} less than 2.00 implies the statistical consistency of the scale transformation, or a good quality of equating.

Results and Discussion

Despite the large amount of missing data, the program converged very smoothly. The program was terminated after 152 UCONs when the maximum logit change reached the level of −.002 and the maximum score residual was .19. The overall item separation reliability was .99, and the person separation reliability was .89. The person ability distribution and item difficulty distribution mirrored the distributions in individual exams. The person and item fit statistics at either exam level or individual level were all satisfactory.

Figure 3 through Figure 10 plot log mean square fit of both persons and items before and after equating for three target exams and S912. For three target exams, fit was almost identical for both items and persons. Though S912 was less consistent, the overall fit patterns were still satisfactory.

Table 1 indicates that distributions of person ability were unaffected by the equating. Standard deviations of ability estimates on the bank scale were shrinking from 1989 to 1992.

Obviously, the shrinkage was not the effect of equating, because it was the reality before the equating.

Table 2 summarizes the performances of item sebsets. For all exams and subsets of items within exams, RMS_{z_i} was smaller than 2.00, indicating none of the exams or subsets of items had problematic behavior. Yet, subsets with a small number of items, such as Subset 1 for A891, Subset 1 for B911, Subset 1 for C921, and Subset 1 for S912, tended to have larger standard deviations of linking shift $\overline{a}$; larger $\overline{D}$, the variation from the whole exam; larger $\overline{z}$; and larger RMS_{z_i}, the accumulated variation. Again, scale transition for S912 was less stable compared with the target exams.

Differences in readiness for the exam possibly contributed to S912's less consistent performance. All 199 S912 subjects were already beginning physicians. Their readiness for Part I or II items was different from that of students who just finished prelinical or clinical education. In the equating, S912 items were calibrated by S912 subjects together with larger groups of Part I or II students. This sampling difference might result in the slight variation of estimation before and after equating.

The results of this analysis suggests that Rasch measurement one-step equating is a valid, efficient and accurate way to construct a measure for longitudinal medical achievement studies. This study further demonstrated that one-step equating technique accomplished by BIGSTEPS could provide good person ability estimates on the whole exam and consistent difficulty estimates for items, even no individual student took the whole hypothesized "exam" and no item was anwered by all students. High quality estimates ensured the quality of equating. This study also suggests that a better control of the number of overlapped items and sampling of subjects may increase the quality of one-step equating.

References

Becker, D. F., & Forsyth, R. A. (1993). An Empirical Investigation of Thurstone and IRT Methods of Scaling Achievement Tests. Journal of Educational Measurement. 29(4), 341-354.

Gonnella, J. S., Hojat, M., Erdmann, J. B., and Veloski, J. J. A case of Mistaken Identity: Signal and Noise in Connecting Performance Assessments Before and After Graduation from Medical School. Academic Medicine. 68(1993):S9-S16.

Lee, O. K., Calibration Matrices for Test Equating. Rasch Measurement SIG Newsletter. 6:1, 202-203.

Levit, E. J., Use of the National Board "Minitest" for Evaluation of Curriculum Change. Journal of Medical Education. 42:930-934, 1967.

McGuire, C. Perspectives in Assessment. Academic Medicine. 68(1993):S3-S8.

Schultz, E. M., Shen, L. and Wright, B. D., An Equal-Interval Scale for Studying Reading Growth. Paper presented at the AERA annual meeting, Boston, 1990.

Verwijnen, M. et al. (1990). A Comparison of an Innovative Medical School with Traditional Schools: An analysis in the Cognitive Domain. In Nooman, Z. M., Schmidt, H. G., & Ezzat, E. S. (Eds.) Innovation in Medical Education: An Evaluation of Its Present Status (pp. 40-49). New York: Springer.

Willoughby, T. L. et al. Edumetric Validity of the Quarterly Profile Examination. Educational and Psychological Measurement. 38:1057-1061, 1978.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. Journal of Educational Measurement, 23, 299-325.

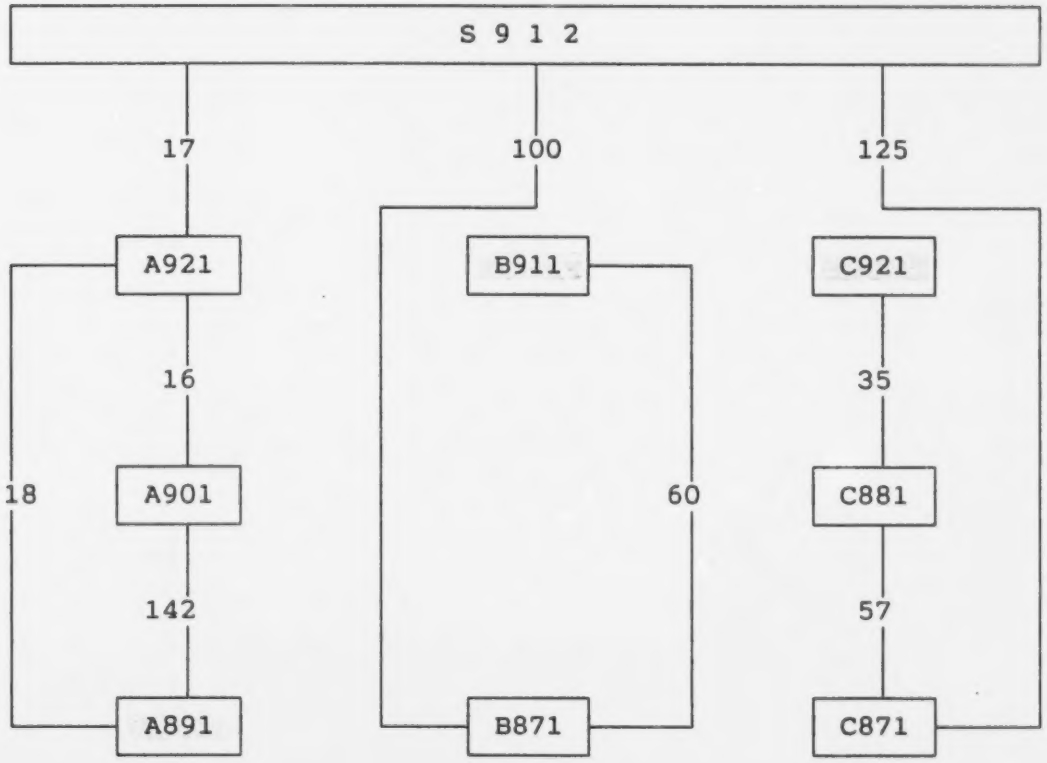

Fig. 1. Linking structure. Numbers between examinations are the numbers of common items shared by two examinations. The highlighted examinations are the target examinations to be equated.

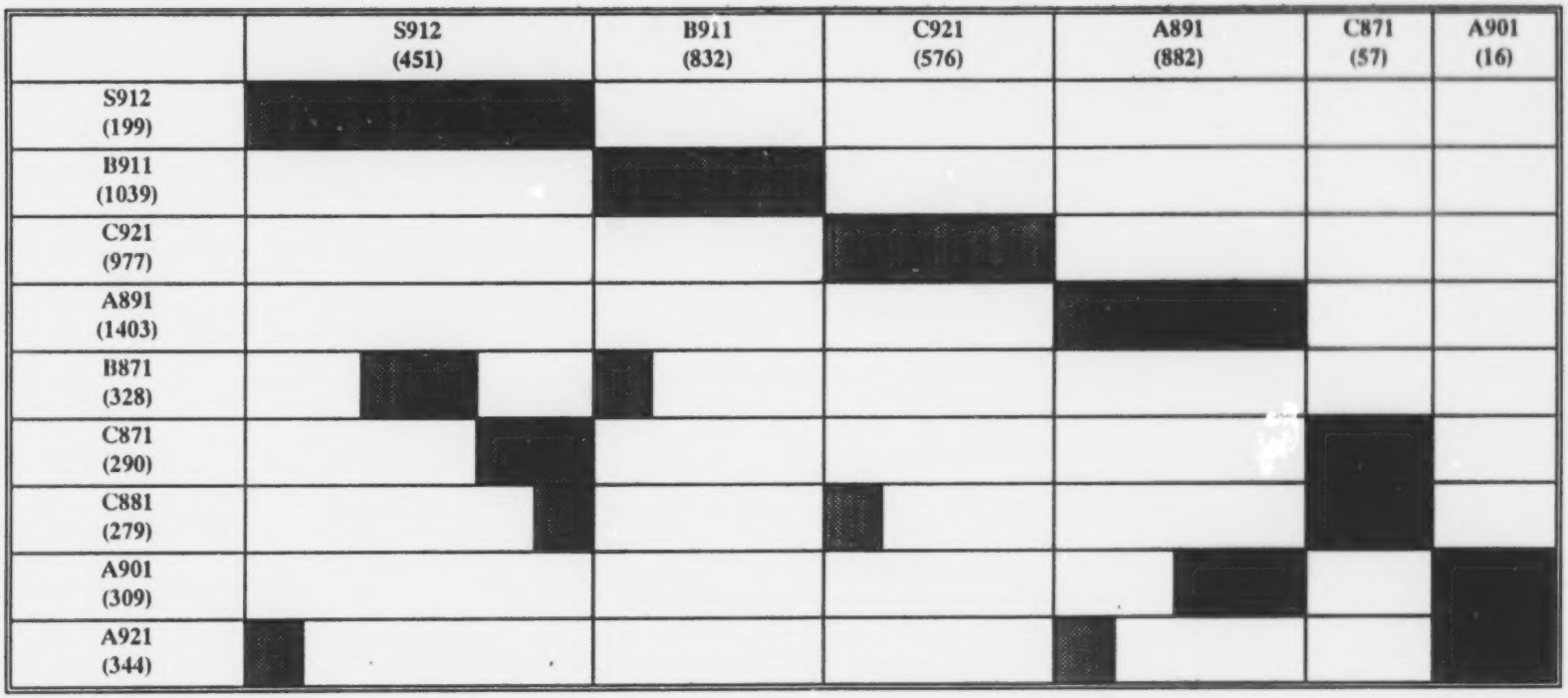

Fig. 2. The data matrix for the Rasch one-step equating. Numbers in parentheses in the first row identify the numbers of items of each exam selected for the equating. Numbers in parentheses in the first column identify the numbers of persons of each exam selected for the equating. In total, 2814 items and 5168 persons were included.

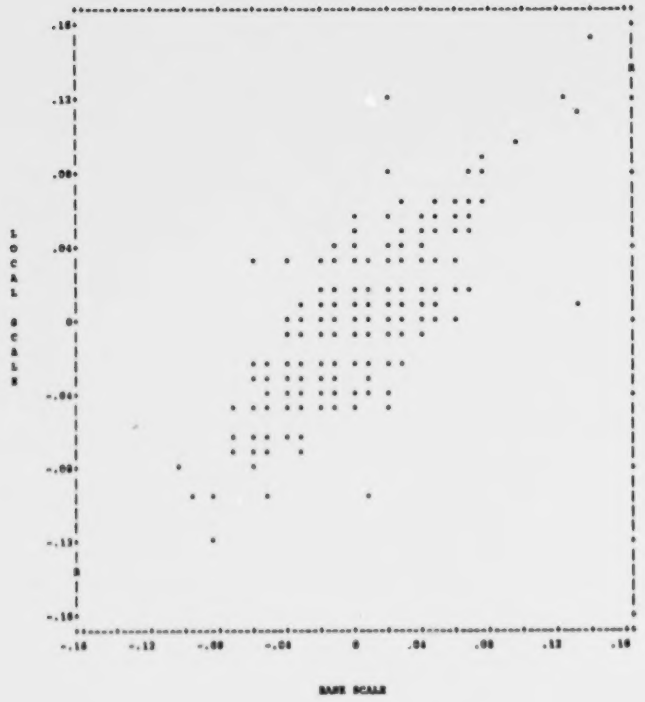

Fig. 3. Plot of log item MMSQ_infit before and after equating for 451 8912 items. Regression statistics: Correlation(S.E.) .92 (.01), Intercept(S.E.) -.001(.001), Slope(S.E.). 86(.020).

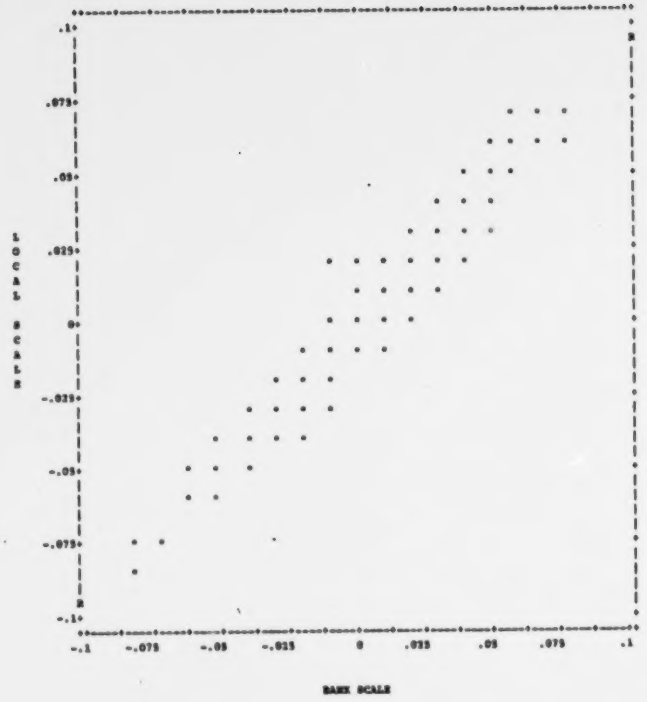

Fig. 4. Plot of log item MMSQ_infit before and after equating for 631 8911 items. Regression statistics: Correlation (S.E.) .97 (.006), Intercept(S.E.) -.0004(.0002), Slope(S.E.) .95(.009)

BEST COPY AVAILABLE

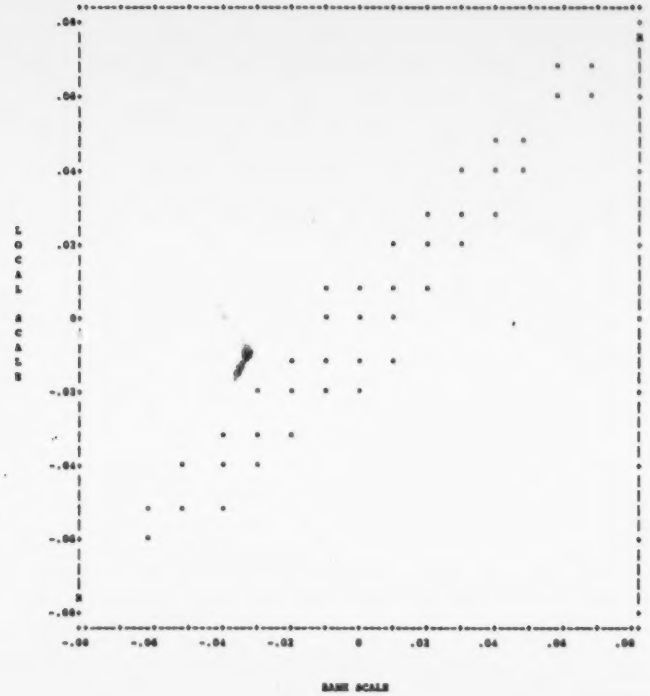

Fig. 5. Plot of log item MNSQ_infit before and after equating for 576 C921 items. Regression statistics: Correlation(S.E.) .97 (.005), Intercept(S.E.) .0003 (.0003), Slope(S.E.) .99(.01)

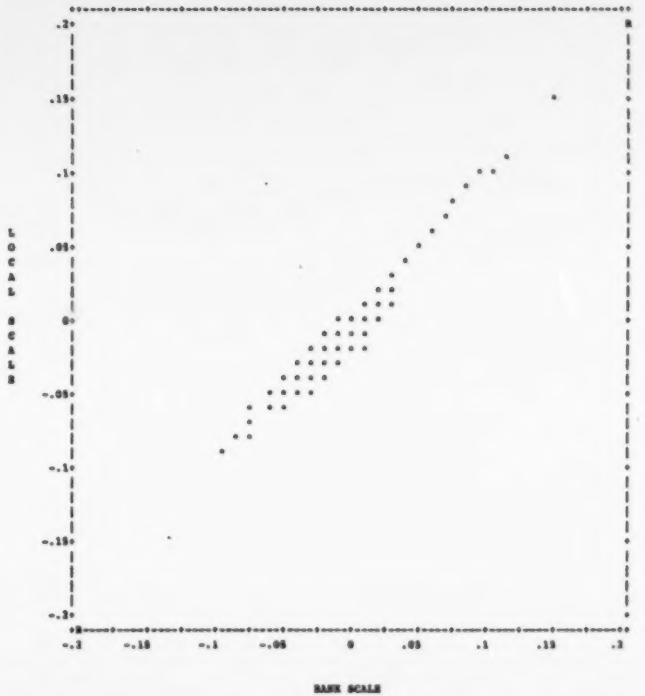

Fig. 6. Plot of log item MNSQ_infit before and after equating for 882 A691 items. Regression statistics: Correlation(S.E.) .99 (<.004), Intercept(S.E.) -.0008(.0001), Slope(S.E.) 1.00(.004)

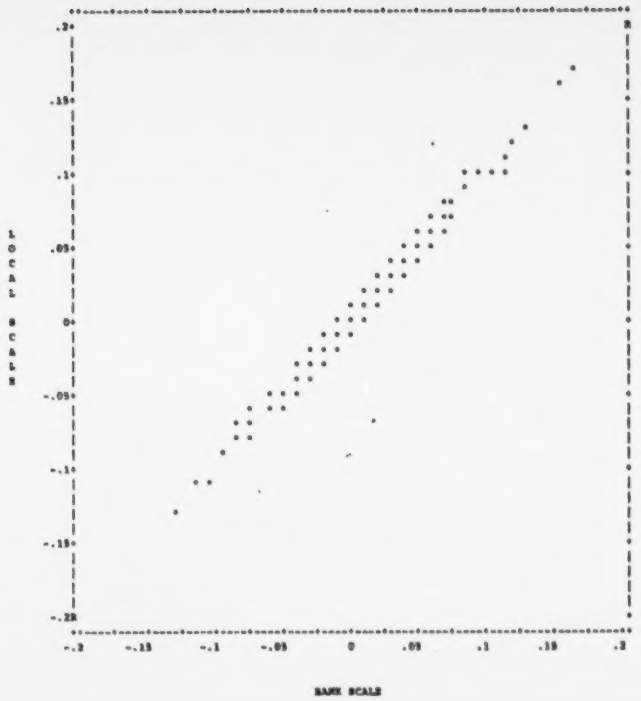

Fig. 7. Plot of log person MNSQ_infit before and after equating for 1403 A001 students. Regression statistics: Correlation(S.E.) .997 (.003), Intercept(S.E.) -.00003(.00007), Slope(S.E.) .999(.002)

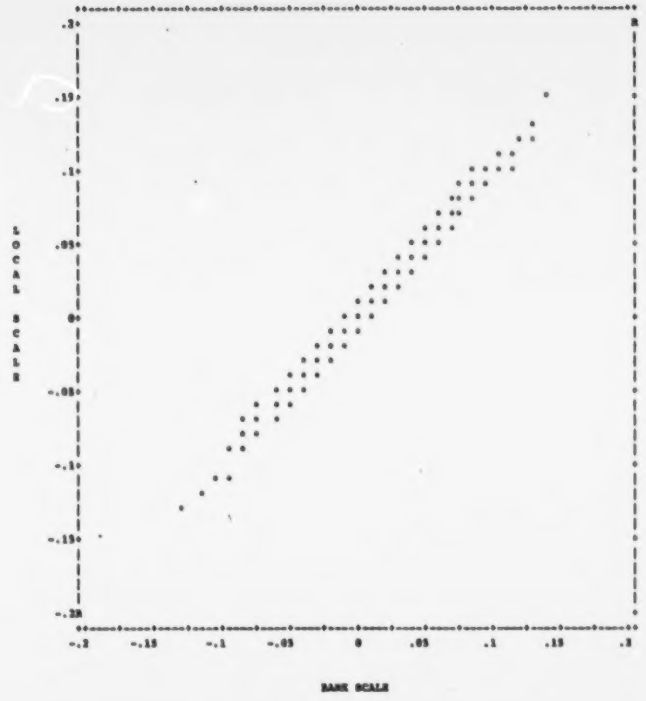

Fig. 8. Plot of log person MNSQ_infit before and after equating for 1039 B011 students. Regression statistics: Correlation(S.E.) .99 (.003), Intercept(S.E.) .0003(.00027), Slope(S.E.).99(.004)

BEST COPY AVAILABLE

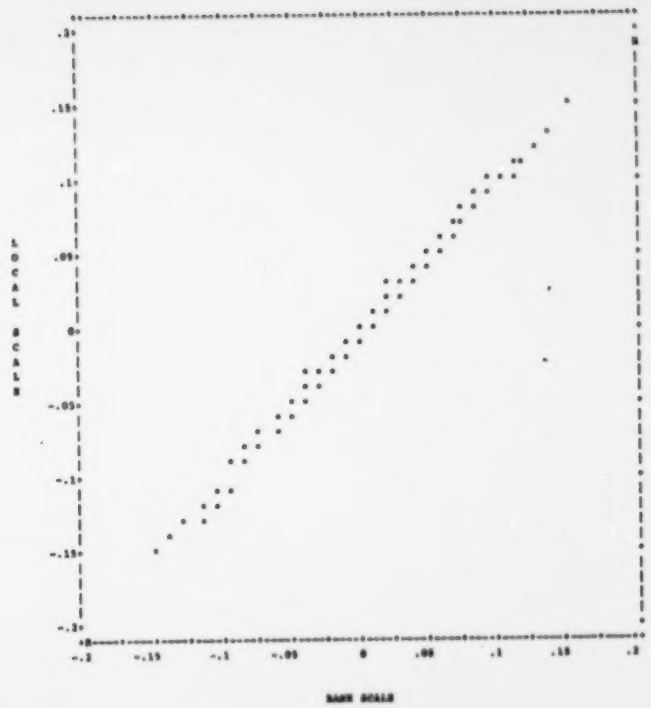

Fig. 9. Plot of log person MSSQ_infit before and after equating for 977 C921 students.
Regression statistics: Correlation(S.E.) .998 (.006), Intercept(S.E.) -.004(.0003)
Slope(S.E.) .99(.003)

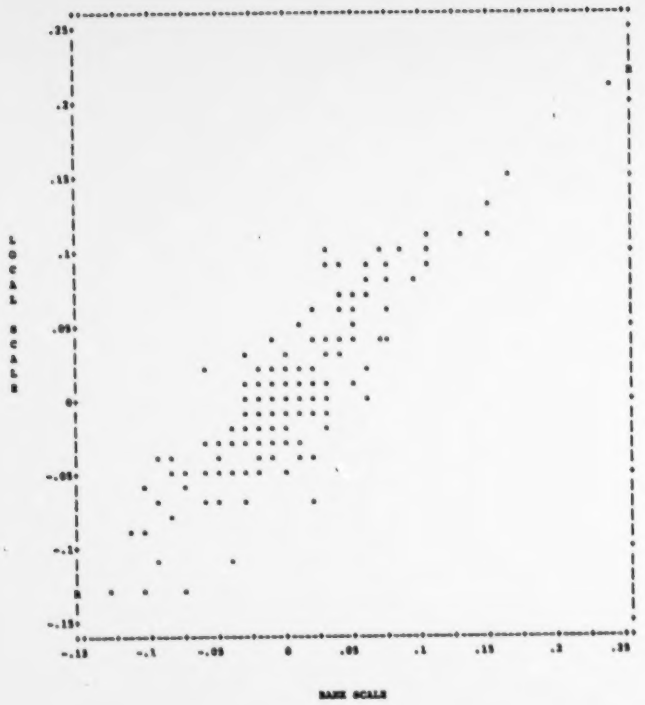

Fig. 10. Plot of log person MSSQ_infit before and after equating for 199 S912 students.
Regression statistics: Correlation(S.E.) .88 (.026), Intercept(S.E.) .00009(.001),
Slope(S.E.) .87(.04)

Table 1

Person Ability Distribution Comparisons and Scale Transformation

	Distribution						Scale Transformation			
	N	Mean	SD	Range	Skewness	Kurtosis	$\bar{d}$	SD	RMS_z	Correlation
						A891				
Bank	1403	.58	.44	3.25	.52	.60	.101	.002	.02	.999
Local	1403	.48	.44	3.25	.52	.60				
						B911				
Bank	1309	.559	.35	2.44	.39	.20	.130	.009	.08	.999
Local	1309	.429	.35	2.43	.39	.20				
						C921				
Bank	977	.697	.317	2.49	.447	.42	-.142	.004	.03	.997
Local	977	.839	.316	2.49	.435	.41				
						S912				
Bank	199	.426	.37	2.24	.225	.32	-.478	.060	.43	.985
Local	199	.904	.38	2.18	.164	.20				

Table 2

Scale Transformation by Subset of Items

	N	$\bar{a}$	SD	Correlation	$\bar{b}$	$\bar{z}$	RMS_{z_i}
				A891			
Total	882	.101	.02				.31
Subset 1	18	.129	.06	.99	-.028	.36	.72
Subset 2	722	.101	.02	.99	.000	.00	.28
Subset 3	142	.097	.03	.99	.004	-.05	.33
				B911			
Total	832	.130	.05	.99			.54
Subset 1	60	.134	.15	.99	-.004	.03	1.65
Subset 2	772	.130	.03	.99	.000	.00	.33
				C921			
Total	576	-.145	.05	.99			.45
Subset 1	35	-.149	.09	.99	.004	-.037	1.04
Subset 2	541	-.144	.05	.99	.001	.022	.38
				S912			
Total	451	-.472	.31	.97			1.33
Subset 1	17	-.555	.27	.97	.083	-.41	1.26
Subset 2	100	-.447	.33	.94	.025	.11	1.58
Subset 3	125	-.475	.46	.94	-.003	-.05	1.83
Subset 4	209	-.476	.15	.99	-.004	-.03	.66